

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/129174>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Machine Learning with Abstention for Automated Liver Disease Diagnosis

Kanza Hamid

PIEAS Biomedical Informatics Laboratory
Pakistan Institute of Engineering and Applied Sciences
Islamabad, Pakistan
kanzahamid@hotmail.com

Amina Asif

PIEAS Biomedical Informatics Laboratory
Pakistan Institute of Engineering and Applied Sciences
Islamabad, Pakistan
a.asif.shah01@gmail.com

Wajid Arshad. Abbasi

PIEAS Biomedical Informatics Laboratory
Pakistan Institute of Engineering and Applied Sciences
Islamabad, Pakistan
wajidarshad@gmail.com

Durre Sabih

Multan Institute of Nuclear Medicine and Radiotherapy
Multan, Pakistan.
dsabih@yahoo.com

Fayyaz-ul-Amir Afsar Minhas

PIEAS Biomedical Informatics Laboratory
Pakistan Institute of Engineering and Applied Sciences
Islamabad, Pakistan
fayyazafsar@gmail.com

Abstract—This paper presents a novel approach for detection of liver abnormalities in an automated manner using ultrasound images. For this purpose, we have implemented a machine learning model that can, not only generate labels (normal and abnormal) for a given ultrasound image but, it can also detect when its prediction is likely to be incorrect. The proposed model abstains from generating the label of a test example if it is not confident about its prediction. Such behavior is commonly practiced by medical doctors who, when given insufficient information or a difficult case, can choose to carry out further clinical or diagnostic tests before generating a diagnosis. However, existing machine learning models are designed in a way to always generate a label for a given example even when the confidence of their prediction is low. We have proposed a novel stochastic gradient descent based solver for the learning with abstention paradigm and use it to make a practical, state of the art method for liver disease classification. The proposed method has been benchmarked on a data set of approximately 100 patients from MINAR, Multan, Pakistan and our results show that the performance of the proposed scheme is at par with medical experts.

Keywords- *Ultrasound, Liver disease, learning with abstention, learning with rejection, machine learning, fatty liver disease, heterogeneous liver texture.*

I. INTRODUCTION

Liver diseases are a cause of major health problems and mortality especially in developing countries such as Pakistan [1]. Fatty liver disease (FLD) and heterogeneous liver texture are among the precursors of more serious liver disorders such as cirrhosis [2]. In FLD, lipid cells start accumulating in the liver whereas heterogeneous liver texture is a consequence of the formation of irregular cells. The detection of these liver disorders can be difficult, especially in their initial stages [3][4].

If these conditions are not detected and treated in time, they may lead to chronic liver disease and cirrhosis which have severe health implications [5].

The most accurate method for diagnosis of such liver diseases is liver biopsy which is invasive, risky, painful and expensive [6]. Non-invasive methods for liver disease diagnosis include ultrasound (US), computed tomography (CT), elastography, etc. These methods are painless and less expensive but are also less accurate than liver biopsy [7]. The use of these diagnostic methods requires access to well-trained medical experts and diagnostic facilities. Automated diagnosis systems for liver disorders can save time and money by acting as a pre-screening service to refer only those individuals for further testing or medical advice who have a high predicted likelihood of a liver disorder.

A number of researchers have implemented different machine learning methods to detect liver abnormalities in an automated fashion. Most of such techniques are primarily based on texture analysis of ultrasound images using statistical features followed by a machine learning classifier such as a Support Vector Machine, Random Forest or hierarchical classification, etc. [8]–[13]. Ultrasound is widely used due to its lower cost and easy availability in comparison to other more sophisticated imaging modalities such as CT or electrography. Wun et al. [8] selected statistical features such as mean, standard deviation, gray level difference, run-length percentage, entropy, etc. for ultrasound characterization and reported an agreement of 89.90% with expert classification. Badawi et al. [9] used a fuzzy logic based model for tissue characterization of liver ultrasound images. They reported specificity and sensitivity values of 92% and 96%, respectively, for fatty liver classification. Yoshida et al. [10]

used multiscale texture analysis for classification of liver ultrasound images with the area under a receiver operating characteristic curve (AUC ROC) of 92%. İçer et al. [11] proposed a method based on the evaluation of liver enzymes with the quantitative grading of fatty liver using ultrasound images. They reported AUC ROC scores of 97.5%, 95.8%, and 94.9% for normal, grade I and grade II fatty liver ultrasound images, respectively. Andrade et al. [12] applied stepwise regression as a feature selection method with the k-nearest neighbor, support vector machine and artificial neural network classifiers for detection of liver steatosis using ultrasound images and reported an accuracy of 79.8%. Minhas et al. [13] proposed a wavelet transform based technique for completely automated classification of normal, heterogeneous and fatty liver disorders with an accuracy of 95%. Owjimehr et al. [14] improved upon approach by using a hierarchical classifier with an accuracy of 97.9%.

In this work, we have identified a major issue with all existing automated diagnosis methods in this domain. All existing ultrasound based liver disease diagnosis systems are designed to always generate a label for an input example even if the predicted label is highly likely to be incorrect. In contrast to existing automated techniques, a medical doctor can either choose to diagnose a patient based on available current information about the patient or alternatively, refrain from generating any decision if the available information is not sufficient to reach a reliable diagnosis. In such cases, a doctor will typically request further diagnostic or clinical tests because the cost of a misdiagnosis can be much higher than that resulting from abstention. In the context of liver disorders, an ideal automated ultrasound-based diagnosis system should follow the same pattern, i.e., it should classify an example only if it is highly confident about its prediction and should reject or abstain from classification otherwise. Such a system can function as a more effective pre-screening service in comparison to existing methods by referring only those patients for further medical examination or expensive or invasive tests such as elastography, CT or biopsy for which the classifier has abstained from classification.

With this background, we have developed an automated liver disease diagnosis system that can, not only classify a given liver ultrasound as normal or abnormal but, can also refrain from classification if it is unsure about the correctness of its prediction. Our model is based on a customized implementation of the *learning with rejection* or abstention framework proposed by Cortes et al. [15]. Our experimental results on a dataset comprising of about 100 subjects collected from medical experts in Pakistan shows that the proposed learning with abstention model of automated diagnosis can be very useful in practice.

Another issue with existing approaches is that most of them use data sets that have been annotated by a single medical expert [13], [14]. However, due to the existence of large inter- and intra-expert variability in the diagnosis of fatty liver disease and heterogeneous liver texture [16], the results of these methods cannot be generalized. To counter this, we collected data from multiple experts and compared the performance of our method with these medical experts.

The rest of the paper is organized as follows: Section II gives the details of the proposed method, Section III presents results and discussion. Conclusions and future work are given in Section IV.

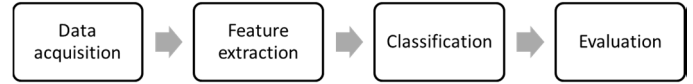


Figure 1- Proposed methodology

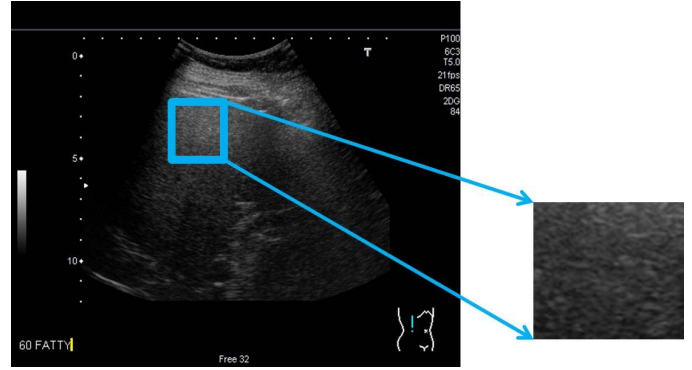


Figure 2- Selected ROI with its ultrasound image

II. METHODS

Our proposed methodology consists of the following steps as shown in Figure 1: ultrasound data acquisition (region of interest (ROIs) selection and annotation by a medical expert), feature extraction, application of various machine learning classification techniques and performance evaluation.

A. Data Acquisition

Our dataset consists of 99 liver ultrasound images. Among these, 43 images are of healthy individuals whereas the remaining 56 have liver abnormalities such as FLD or heterogeneous liver texture. All these images were acquired at Multan Institute of Nuclear Medicine and Radiotherapy (MINAR) Multan, Pakistan, by the author (DS) using a Toshiba Aplio 500 B-mode digital ultrasound machine. The frequency for tissue harmonic imaging was 5 MHz and a convex probe was used. The size of each acquired image is 560×450 pixels and the image was saved as a bitmap file. For these 99 images, 114 64×64 pixel region of interest (ROIs) were selected by the medical expert for annotation into normal or abnormal. All subsequent processing is done on these selected ROIs. An example of a liver ultrasound image with its annotation and ROI is shown in Figure 2. The dataset used here is available through: <http://faculty.pieas.edu.pk/fayyaz/software.html#LWA>. To analyze the performance of the proposed system in comparison to medical experts, we developed an online data collection server and collected annotations for 34 ultrasound images from 12 different experienced medical experts.

B. Feature Extraction

All existing methods use complicated feature extraction techniques. In this work, we chose to use the normalized raw pixel values of the 64×64 ROIs as features. This results in a

4096-dimensional feature vector for a given example. As discussed in the results section, these simple features offer comparable or better accuracy than more sophisticated statistical features.

C. Classification

In order to test our hypothesis that learning with abstention is effective for liver disease diagnosis, we compare the performance of our implementation of *learning with abstention* with conventional classification techniques. Henceforth, we provide details of various classification methods used in this work.

1) Nearest Neighbor (NN)

As a baseline, the nearest neighbor classifier was used to classify data into normal and abnormal classes [17]. Euclidean distance metric was used for distance calculations in the classifier.

2) Support Vector Machine (SVM)

A support vector machine (SVM) finds a maximum margin linear discriminant function $h(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$ to classify the feature representation $\phi(\mathbf{x})$ of an example \mathbf{x} using a weight vector \mathbf{w} and a bias parameter b . An SVM determines the optimal values of \mathbf{w} and b by using a training set $S = \{(\mathbf{x}_i, y_i) | i = 1, 2 \dots N\}$ of examples with corresponding labels $y_i = -1$ or $y_i = +1$ for normal and abnormal cases, respectively. This is done by solving the following optimization problem:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N l_{SVM}(h, \mathbf{x}_i, y_i)$$

Here, the first term $\|\mathbf{w}\|^2$ is responsible for margin maximization and second term controls the number of misclassification over training data by using hinge loss function $l_{SVM}(h, \mathbf{x}_i, y_i) = \max\{0, 1 - y_i h(\mathbf{x}_i)\}$. The hinge loss function penalizes misclassifications and margin violations. The hyper-parameter C is the weighting factor between these two terms and is chosen through cross-validation [18].

3) Learning with Abstention (LWA)

Conventional classifiers are designed to always produce a label given an example which can either be correct or incorrect. As discussed earlier, it would be more practical if a classifier can abstain from generating a label when it is not confident about its decision instead of producing a misclassification. In this work, we have implemented a classifier that can refrain from generating labels for such test examples. The idea of learning with abstention was proposed by Cortes et al. [15]. Such a classifier can generate three different types of labels in our case: normal (-1), abnormal ($+1$) or Reject (R) which corresponds to an abstention from classification. We followed the same principle for construction of the LWA classifier as in Cortes et al [15]. However, unlike their approach, we have solved the optimization problem of the LWA classifier using a stochastic gradient based solver [19].

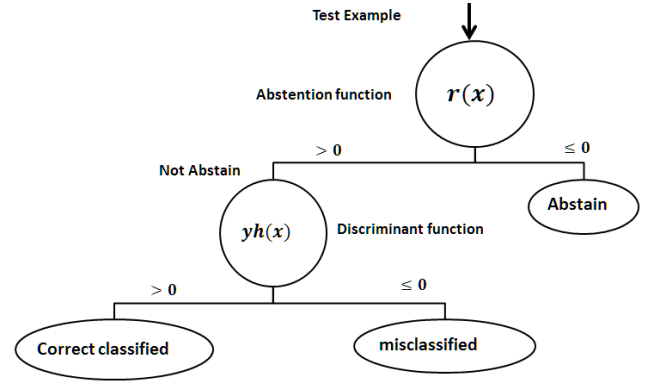


Figure 3- Systematic diagram of proposed method

As discussed in the work by Cortes et al. [15], LWA requires two decision functions: a discriminant function $h(\mathbf{x})$ which is the same as in a standard SVM and an abstention function $r(\mathbf{x}) = \mathbf{u}^T \phi(\mathbf{x}) + b'$ that uses different weight and bias parameters. The objective of LWA is to simultaneously learn both these functions in a way that the rejection function produces a positive score $r(\mathbf{x}) > 0$ only if the discriminant function is expected to correctly classify the given example. If $r(\mathbf{x}) < 0$, the classifier is not confident about the correctness of the label generated by its discriminant function and the example is rejected (abstention). A systematic representation of this concept is shown in Figure 3.

Similar to a conventional SVM, a large margin LWA classifier can be developed through the principle of structural risk minimization [17] by simply using a loss function that takes abstentions into account. For the implementation of the LWA classifier, we use the loss function and its convex over-approximation given by Cortes et al. [15] which works as follows:

- i. *Correct classification without rejection*: In the scenario in which an example is not rejected ($r(\mathbf{x}) > 0$) and is classified correctly ($yh(\mathbf{x}) > 0$), no loss is incurred.
- ii. *Misclassification without rejection*: An example that is not rejected ($r(\mathbf{x}) > 0$) but is misclassified ($yh(\mathbf{x}) \leq 0$), incurs a loss of 1.0.
- iii. *Abstention*: The abstention or rejection of an example ($r(\mathbf{x}) < 0$) incurs a loss of $c \in (0, 0.5)$. The hyper-parameter c is set by the user and it controls the cost and, consequently, the number of rejections. A small c will produce more rejections and vice versa.

The loss function can be written mathematically formulated as follows:

$$l(h, r, \mathbf{x}, y) = \mathbb{I}(yh(\mathbf{x}) \leq 0) \mathbb{I}(r(\mathbf{x}) > 0) + c \mathbb{I}(r(\mathbf{x}) \leq 0) \quad (1)$$

Here $\mathbb{I}(\cdot)$ is the indicator function whose value is 1.0 if its argument is true and 0.0 otherwise. This loss function is non-linear, non-convex and difficult to optimize. Its convex over-approximation can be written as [15]:

$$l_{LWA}(h, r, \mathbf{x}, y) = \max \left(0, 1 + \frac{1}{2} (r(\mathbf{x}) - yh(\mathbf{x})), c(1 - \beta r(\mathbf{x})) \right) \quad (2)$$

Here, $\beta = \frac{1}{1-2c}$. Notice that this loss function will always penalize abstentions ($r(\mathbf{x}) < 0$) and misclassification

($yh(\mathbf{x}) < 0$). For a more detailed description of the loss function, the interested reader is referred to [15].

Following the principle of structural risk minimization and using the above loss function, the LWA learning problem can be expressed as the following optimization problem:

$$\min_{\mathbf{w}, \mathbf{u}, b, b'} J(\mathbf{w}, \mathbf{u}, b, b') = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{\lambda'}{2} \|\mathbf{u}\|^2 + \sum_{i=1}^N l_{LWA}(h, r, \mathbf{x}_i, y_i) \quad (3)$$

Here, the first two terms control the margin for the discriminant and rejection functions using hyper-parameters λ and λ' whereas the second term is responsible for loss-minimization. The solution to this optimization problem will result in optimal values of weights for both decision functions so that both misclassifications and abstentions are minimized. We have developed a stochastic gradient solver for the LWA optimization problem in equation (3). The proposed algorithm is inspired from the Pegasos solver for conventional support vector machines proposed by Shalev-Shwartz et al. [19]. It offers an easier and more scalable alternative to quadratic programming or sequential minimal optimization methods typically used in SVMs. The proposed method is based on step-wise iterative updates to weight parameters in a direction opposite to the sub-gradients of the objective function using a single randomly chosen training example. The weight update equations at iteration t can be written as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \quad (4)$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta' \nabla_{\mathbf{u}} \quad (5)$$

Here, $\eta = \frac{1}{\lambda t}$ and $\eta' = \frac{1}{\lambda' t}$ are the step-sizes for the gradients $\nabla_{\mathbf{w}} = \frac{\partial J}{\partial \mathbf{w}}$ and $\nabla_{\mathbf{u}} = \frac{\partial J}{\partial \mathbf{u}}$, respectively. For a randomly chosen training example \mathbf{x} with label y , the sub-gradients of the objective function can be computed by taking the derivative of the objective function with respect to the weight parameters. Consequently, the sub-gradients can be written as follows:

$$\nabla_{\mathbf{w}} = \begin{cases} \lambda \mathbf{w} - \frac{1}{2} y \phi(\mathbf{x}) & \text{if } 1 + \frac{1}{2} (r(\mathbf{x}) - yh(\mathbf{x})) > \max(0, c(1 - \beta r(\mathbf{x}))) \\ \lambda \mathbf{w} & \text{else} \end{cases}$$

$$\nabla_{\mathbf{u}} = \begin{cases} \lambda' \mathbf{u} + \frac{1}{2} \phi(\mathbf{x}) & \text{if } 1 + \frac{1}{2} (r(\mathbf{x}) - yh(\mathbf{x})) > \max(0, c(1 - \beta r(\mathbf{x}))) \\ \lambda' \mathbf{u} - c \beta \phi(\mathbf{x}) & \text{if } c(1 - \beta r(\mathbf{x})) > \max(0, 1 + \frac{1}{2} (r(\mathbf{x}) - yh(\mathbf{x}))) \\ \lambda' \mathbf{u} & \text{else} \end{cases}$$

Substituting the above sub-gradient calculations into the weight update equations leads us to the complete algorithm for learning with abstention which is given in Figure 4. It is important to note that the bias term has been omitted for clarity and it is trivial to obtain bias update equations. As discussed earlier, the proposed algorithm operates by selecting a training example from the training data uniformly at random and calculating the sub-gradient of the objective function and performing weight updates in the direction opposite to the sub-gradient. The hyper-parameters λ, λ' and c are selected through cross-validation. Once the optimal weight vectors have been obtained, the classifier can generate labels for a given test example: if $r(\mathbf{x}) < 0$, the example is rejected as the classifier is not confident about its prediction, otherwise, the decision function $h(\mathbf{x})$ is used to determine the class (normal or abnormal) for the given example. This algorithm has been

implemented in Python 2.7. and its implementation is available online at the URL: <http://faculty.pieas.edu.pk/fayyaz/software.html#LWA>.

D. Evaluation

For evaluation of the performance of all classifiers used in this work, we have used 5-fold cross-validation. K fold cross-validation [20] is the method of choice for evaluating machine learning problems with small data sets. In this approach, data is divided into K sets, leaving K-1 sets for training, testing is performed on the held-out set and this process is repeated for all sets. As for performance metrics, we have used the area under the receiver operating characteristic curve (AUC-ROC) as well as the number of misclassifications and abstentions. AUC-ROC is obtained by plotting the specificity of the classifier at different decision thresholds vs. its sensitivity. The higher the value of AUC-ROC, the better the classifier [20].

Learning with Abstention Using Stochastic Gradients

INPUT: Training set $S = \{(\mathbf{x}_i, y_i) | i = 1, 2 \dots N\}$

HYPER-PARAMETERS:

Regulatization parameter for \mathbf{w} : $\lambda > 0$

Regulatization parameter for \mathbf{u} : $\lambda' > 0$

Abstention Penalty: $c \in (0, 0.5)$

Number of iterations: $T > 0$

INITIALIZE: $\mathbf{w}_1 = \mathbf{0}$, $\mathbf{u}_1 = \mathbf{0}$, $\beta = 1/(1 - 2c)$

For $t = 1, 2, \dots, T$

Choose example $(\mathbf{x}, y) \in S$ uniformly at random

Calculate $h(\mathbf{x}) = \mathbf{w}_t \phi(\mathbf{x})$

Calculate $r(\mathbf{x}) = \mathbf{u}_t \phi(\mathbf{x})$

If $\left(1 + \frac{1}{2} (r(\mathbf{x}) - yh(\mathbf{x}))\right) > \max(0, c(1 - \beta r(\mathbf{x})))$ then:

$$\mathbf{w}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) \mathbf{w}_t + \frac{1}{2\lambda t} y \phi(\mathbf{x})$$

$$\mathbf{u}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) \mathbf{u}_t - \frac{1}{2\lambda' t} \phi(\mathbf{x})$$

ElseIf $c(1 - \beta r(\mathbf{x})) > \max\left(0, 1 + \frac{1}{2} (r(\mathbf{x}) - yh(\mathbf{x}))\right)$ then:

$$\mathbf{w}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) \mathbf{w}_t$$

$$\mathbf{u}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) \mathbf{u}_t + c \beta \phi(\mathbf{x})$$

Else:

$$\mathbf{w}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) \mathbf{w}_t$$

$$\mathbf{u}_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) \mathbf{u}_t$$

OUTPUT: $\mathbf{w} = \mathbf{w}_{T+1}$, $\mathbf{u} = \mathbf{u}_{T+1}$

Classification with Abstention Using Stochastic Gradients

INPUT: Test example (\mathbf{x}, y)

Calculate $h(\mathbf{x}) = \mathbf{w} \phi(\mathbf{x})$

Calculate $r(\mathbf{x}) = \mathbf{u} \phi(\mathbf{x})$

If $r(\mathbf{x}) < 0$:

Output "Reject"

Else:

Output $h(\mathbf{x})$

Figure 4- Pseudo code of proposed classifier

III. RESULTS AND DISCUSSION

A. Comparison of LWA with conventional classification

Figure 5 shows the results of different classification techniques in terms of AUC-ROC vs. the fraction of abstentions. The nearest neighbor classifier gives AUC-ROC of 78% while, conventional SVM performs significantly better with the AUC-ROC of 87%.

In order to compare the performance of the proposed LWA classifier, we refer to Figure 5 in which we plot the AUC-ROC of our proposed scheme with and without taking examples that the model has abstained from classifying. The AUC-ROC of NN and SVM are also plotted as a reference. It can be noticed that the AUC-ROC of LWA on accepted (not abstained) examples is always better and its overall AUC-ROC is comparable to the AUC-ROC of conventional SVM. As discussed earlier, the increase in abstention penalty decreases the fraction of abstentions. For the low value of $c = 0.1$, the LWA classifier rejects all examples whereas for high value of $c = 0.5$, no abstentions take place. Furthermore, As expected, when the fraction of abstention drops to zero for large values of c , the performance of LWA becomes comparable to a conventional SVM. However, for $c = 0.12$, the fraction of abstention is equal to 53% with an AUC-ROC of 95. For $c = 0.17$, the fraction of abstention is equal to 21% with an AUC-ROC of 93 and when $c = 0.3$, the fraction of abstention is equal to 4% with an AUC-ROC of 91. This shows that the LWA classifier achieves near perfect classification AUC-ROC if it is permitted to abstain from producing labels for 53% test examples. LWA has automatically detected that its confidence for correctly predicting these examples is low and thus abstained from these misclassifications. This shows the effectiveness of the proposed approach in comparison to conventional classification techniques. The python implementation of the LWA classifier runs in under 5-6 minutes on a laptop with an Intel core i5-3317U 1.70 GHz processor and 4 GB RAM.

B. Re-Evaluation of rejected examples by medical expert

The 7 test cases from 53% of data for which the LWA classifier abstained from generating labels were given to an experienced radiologist (DS) for re-evaluation. The radiologist was not provided the original labels for these cases and was asked to diagnose these cases. It is interesting to notice that, for 3 out of these 7 cases, the radiologist generated labels were different from the original labels. These cases are shown in Figure 6. This shows that the abstentions produced by the proposed LWA method were indeed difficult to classify even for trained medical experts. These cases can refer to further testing through elastography, CT or biopsy. These results clearly indicate the effectiveness of the proposed approach.

C. Comparison with medical expert's analysis

Since the previous state of the art methods in this domain used annotations by a single medical expert which are inherently subjective due to intra and inter-observer variability, we compare the performance of our method with that of trained radiologists and not with previous methods.

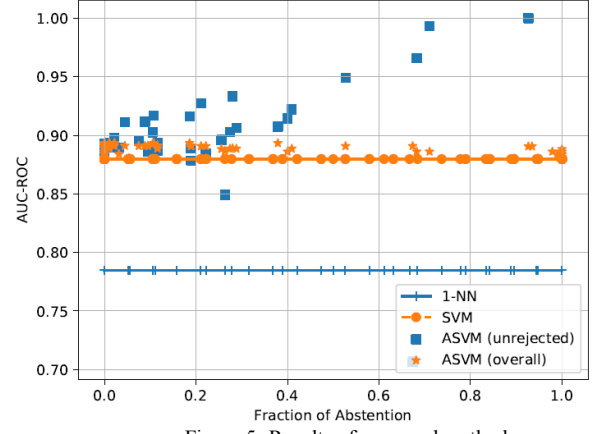


Figure 5- Results of proposed method

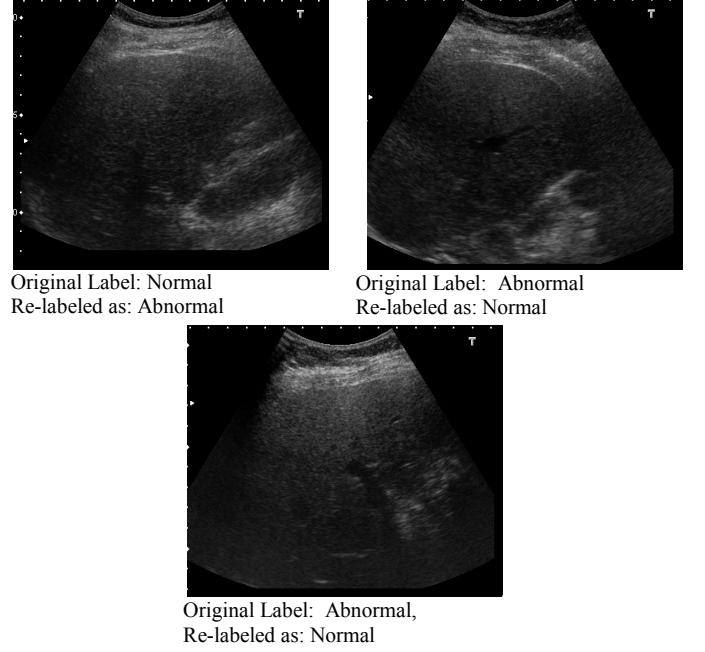


Figure 6- Results of re-labeling by a medical expert of the LWA-rejected images

For this purpose, we analyzed annotations from 12 experienced medical experts over 34 different ultrasound images [21]. We evaluated the degree of correspondence of the decisions of each medical expert with the consensus of all medical experts by computing the AUC-ROC of the labels generated by a particular expert with the consensus label of all radiologists. Since different radiologists annotated a different number of images, we first plot the AUC-ROC of each doctor against the fraction (out of 34) of images annotated by that doctor (Figure 7). The weighted average of the AUC-ROC for all experts is below 80% with the highest AUC-ROC of 89.5% for all annotations. Figure 8 plots the performance of medical experts in comparison to the automated techniques discussed in this work. This shows that the proposed system can perform on par with medical experts.

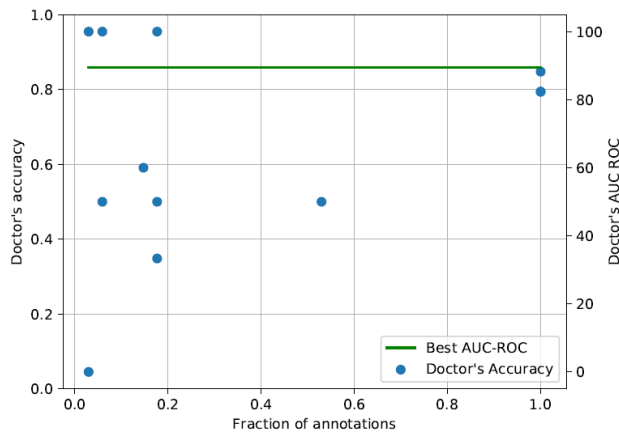


Figure 7 Analysis of annotations by 12 medical experts

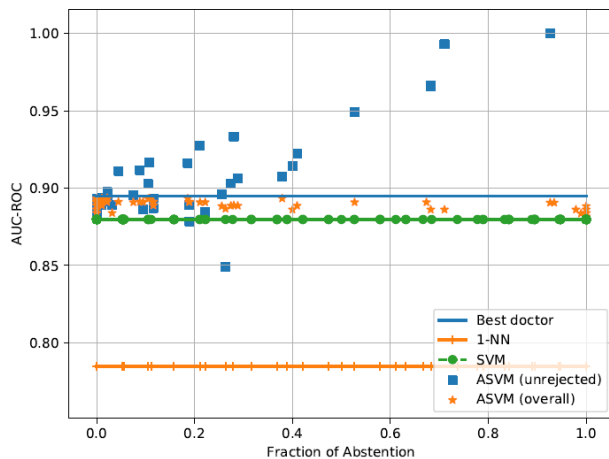


Figure 8 Comparison between doctor's AUC-ROC and classifiers

IV. CONCLUSIONS AND FUTURE WORK

In this work, we have developed a novel approach for classification of liver ultrasound images into normal and abnormal cases. The novelty of our approach lies in using a *learning with abstention* model for classification. Our proposed method is able to automatically identify cases for which it does not have high enough confidence of generating accurate predictions and identify outliers. Thus, the model can be thought of an artificial intelligence (AI) system that *knows what it doesn't know*. Our results clearly show that the proposed system is very useful in a practical setting and can help both patients and medical doctors by saving their time, money and the inconvenience of undergoing painful or expensive tests. We have also proposed a novel stochastic gradient based solver for the LWA framework. The proposed scheme can be applied in other domains as well. In future, we aim to extend this method to multi-class classification and evaluate our performance on a large independent test set with elastography data. We also plan to build a publicly accessible web server implementation of our method.

Author contributions: **KH** – Implementation, testing and manuscript writing, **AA** – Mathematical formulation and prototype

development, **WA** – Data collection server development, **DS** – Medical data collection, annotation and supervision, **FuAAM** – Original idea, supervision and manuscript writing.

Acknowledgements: KH and AA are supported by MS and Ph.D. scholarships, respectively, from the Information Technology and Telecom Endowment Fund, PIEAS. WA is supported by a Ph.D. grant from the higher education commission under the 5000 indigenous Ph.D. scholars scheme.

REFERENCES

- [1] T. S. Khan, F. Rizvi, and A. Rashid, "Hepatitis C seropositivity among chronic liver disease patients in Hazara, Pakistan," *J Ayub Med Coll Abbottabad*, vol. 15, no. 2, pp. 53–5, 2003.
- [2] G. C. Farrell and C. Z. Larter, "Nonalcoholic fatty liver disease: from steatosis to cirrhosis," *Hepatology*, vol. 43, no. S1, 2006.
- [3] M. Carmiel-Haggai, A. I. Cederbaum, and N. Nieto, "A high-fat diet leads to the progression of non-alcoholic fatty liver disease in obese rats," *FASEB J.*, vol. 19, no. 1, pp. 136–138, 2005.
- [4] S. Sell, "Heterogeneity and plasticity of hepatocyte lineage cells," *Hepatology*, vol. 33, no. 3, pp. 738–750, 2001.
- [5] S. Dam-Larsen *et al.*, "Long term prognosis of fatty liver: risk of chronic liver disease and death," *Gut*, vol. 53, no. 5, pp. 750–755, 2004.
- [6] R. Kutcher *et al.*, "Comparison of sonograms and liver histologic findings in patients with chronic hepatitis C virus infection," *J. Ultrasound Med.*, vol. 17, no. 5, pp. 321–325, 1998.
- [7] L. Castera, X. Forns, and A. Alberti, "Non-invasive evaluation of liver fibrosis using transient elastography," *J. Hepatol.*, vol. 48, no. 5, pp. 835–847, 2008.
- [8] Y. Wun and R. Chung, "Ultrasound characterization by stable statistical patterns," *Comput. Methods Programs Biomed.*, vol. 55, no. 2, pp. 117–126, 1998.
- [9] A. M. Badawi, A. S. Derbala, and A.-B. M. Youssef, "Fuzzy logic algorithm for quantitative tissue characterization of diffuse liver diseases from ultrasound images," *Int. J. Med. Inf.*, vol. 55, no. 2, pp. 135–147, 1999.
- [10] H. Yoshida, D. D. Casalino, B. Keserci, A. Coskun, O. Ozturk, and A. Savranlar, "Wavelet-packet-based texture analysis for differentiation between benign and malignant liver tumours in ultrasound images," *Phys. Med. Biol.*, vol. 48, no. 22, p. 3735, 2003.
- [11] S. İcer, A. Coşkun, and T. İkizceli, "Quantitative grading using grey relational analysis on ultrasonographic images of a fatty liver," *J. Med. Syst.*, pp. 1–8, 2012.
- [12] A. Andrade, J. Silva, J. Santos, and P. Belo-Soares, *Classifier approaches for liver steatosis using ultrasound images. Procedia Technol* 2012; 5: 763–70.
- [13] F. ul A. A. Minhas, D. Sabih, and M. Hussain, "Automated Classification of Liver Disorders using Ultrasound Images," *J. Med. Syst.*, vol. 36, no. 5, pp. 3163–3172, Oct. 2012.
- [14] M. Owjimehr, H. Danyali, and M. S. Helfroush, "An Improved Method for Liver Diseases Detection by Ultrasound Image Analysis," *J. Med. Signals Sens.*, vol. 5, no. 1, pp. 21–29, 2015.
- [15] C. Cortes, G. DeSalvo, and M. Mohri, "Learning with Rejection," in *Algorithmic Learning Theory*, 2016, pp. 67–82.
- [16] S. Strauss, E. Gavish, P. Gottlieb, and L. Katsnelson, "Interobserver and intraobserver variability in the sonographic assessment of fatty liver," *Am. J. Roentgenol.*, vol. 189, no. 6, pp. W320–W323, 2007.
- [17] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.
- [18] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology," *PLoS Comput. Biol.*, vol. 4, no. 10, p. e1000173, 2008.
- [19] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for SVM," *Math. Program.*, vol. 127, no. 1, pp. 3–30, 2011.
- [20] E. Alpaydin, *Introduction to machine learning*. MIT Press, 2014.
- [21] "hepaticus server," *Hepaticus*.